**МИНИСТЕРСТВО ОБРАЗОВАНИЯ И НАУКИ РОССИЙСКОЙ ФЕДЕРАЦИИ**

**ФЕДЕРАЛЬНОЕ ГОСУДАРСТВЕННОЕ БЮДЖЕТНОЕ ОБРАЗОВАТЕЛЬНОЕ УЧРЕЖДЕНИЕ ВЫСШЕГО ОБРАЗОВАНИЯ**

**«Рязанский государственный радиотехнический университет»**

**(ФГБОУ ВО «РГРТУ», РГРТУ)**

«Утверждаю»

Заведующий кафедрой

ВПМ Пылькин А.Н.

\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_  *(подпись)*

«\_\_\_» \_\_\_\_\_\_\_\_\_\_ 2017 г.

ОТЧЕТ

О НАУЧНОЙ РАБОТЕ СТУДЕНТА

**4 курса, 344 группы**

*(Яровова О.С.)*

Рязань 2017

Направление подготовки: 09.03.03 «Прикладная информатика»

*(шифр и наименование)*

Тема НИР: «Исследование методов представления цепочек преобразования типов формообразования естественного языка с помощью графов»

|  |  |  |
| --- | --- | --- |
| № | Результаты научной работы студента | Результаты |
| 1 | Рассмотрение сферы исследования и обоснование её актуальности | Составлен отчет об актуальности представления цепочек преобразования типов формообразования естественного языка с помощью графов |
| 2 | Формирование классификации типов формообразования естественного языка | Сформирована классификация типов формообразования естественного языка для дальнейшего представления их с помощью графов |
| 3 | Изучение и анализ методов генерации и определения | Проведён анализ методов генерации о определения |
| 4 | Отчет о полученных результатах | Составлен отчет по результатам работы |

Студент \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_(Яровова О.С.)

*(Ф.И.О. студента, подпись)*

«\_\_\_» \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_2017 г.

Научный руководитель \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_(Пруцков А.В.)

*(Ф.И.О., подпись)*

*Студент четвертого курса Яровова О.С.*

*Направление 09.03.03. «Прикладная информатика»*

**Отчет об актуальности сферы исследования**

Во все времена языковая грамотность является неотъемлемой частью любого образованного человека. Каждый из существующих языков обладает своими уникальными чертами и правилами, которые подчас вызывают сомнения в том, какими грамматическими характеристиками обладает слово, какую форму того или иного слова нужно употребить в определенном контексте и как получить необходимую форму. Для решения описанных задач используются информационные системы, удобные и распространенные, которые содержат необходимую информацию.

Русская грамматика и словообразование являются одними из самых сложных не зависимо от того является ли этот язык родным для изучающего или нет. Именно поэтому в настоящей работе рассматривается формообразование глаголов русского. Выбор данной части речи обусловлен многообразием ее форм и сложностей, связанных с ними, ведь правила русского формообразования глаголов сложны, разнообразны и богаты на исключения, именно поэтому для более эффективного изучения необходимо представление этапов формообразования необходимо осуществить наиболее наглядно.

Реализация этой задачи возможна при помощи представления преобразований в виде дерева цепочек преобразований. Для его реализации необходимо осуществить классификацию глаголов по типам формообразования и описать получение цепочек словоформ последовательностью преобразований. После этого необходимо оптимизировать полученные цепочки, путем построения древовидной структуры, что позволит сократить скорость определения грамматического значения и основы. Полученные структуры дадут возможность программного получения из любой введенной формы глагола его грамматических характеристик и инфинитива, от которого образована рассматриваемая форма, также возможны и обратные преобразования - получение формы с необходимыми грамматическими характеристиками по инфинитиву.

Справочная система по морфологии русского языка, на основе рассмотренного подхода, позволит:

получить грамматические характеристики введенной глагольной формы;

получить инфинитив глагола по одной из его форм;

образовывать форму глагола в нужном виде, времени, лице, числе;

образовывать особые формы глагола (причастия, деепричастия) в нужном виде, времени, лице, числе, падеже;

Данная система может стать полезной для многих, чья жизнь в той или иной мере связана с русским языком. Школьникам она будет полезна для решения вопросов, связанных с трудностями, возникающими при выполнении домашних заданий: поиске инфинитива, определению грамматических характеристик слова. Иностранные студенты здесь смогут найти необходимую информацию, которая поможет им определить правильность употребления той или иной формы в построенной ими фразе на русском языке. Да и каждый человек сможет просто проверить себя, найдя интересующую его информацию.

Также возможна реализация метода и для других естественных языков. Справочная система может стать частью более крупных информационных ресурсов для автоматизации обучения и проверки знаний.

*Студент четвертого курса Яровова О.С.*

*Направление 09.03.03. «Прикладная информатика»*

**Классификация глаголов по типам формообразования**

Глагол – часть речи русского языка, обозначающая действия или состояния разного характера.

Глагол обладает различными грамматическими категориями и соответствующими словоформами. Как и другие части речи он имеет категории числа, рода и лица. Однако есть и уникальные категории, присущие только глаголу, например, категории вида, залога, наклонения и времени.

Несомненно, глагол – самая сложная с точки зрения формообразования часть речи русского языка, так как обладает уникальными признаками, присущими лишь глаголу, поэтому была проведена классификация именно этой части речи, в результате которой были составлены цепочки преобразования, соответствующие универсальному методу генерации

Для достижения поставленной цели классификации глаголов русского языка необходимо выделить (составить) полную парадигму, описывающую все словоформы для каждого рассматриваемого инфинитива.

Процесс классификации можно условно разделить на несколько этапов. Сначала необходимо выделить все словоформы рассматриваемых слов, получив их с помощью элементарных преобразований. Далее необходимо разделить полученные цепочки словоформ на типы (группы), обязательным условием, выполняемым в каждой группе, должно быть полное совпадение алгоритмов для каждой цепочки в целом, так и для каждой соответствующей словоформы в отдельности. Полученные группы и являются классификациями глаголов по типам. Глаголы, которые нельзя отнести ни к одному из типов и изменяемые индивидуальным образом, выделяются в отдельную группу.

Таким образом в результате проведенной классификации было рассмотрено 2500 инфинитивов, которые разделились на 124 типа, к глаголам, изменяемым индивидуальным образом отнесено 134.

*Студент четвертого курса Яровова О.С.*

*Направление 09.03.03. «Прикладная информатика»*

**Изучение и анализ метода генерации и определения**

Способ представления формообразования и алгоритмы генерации и определения форм слов взаимосвязаны между собой.

Задачей алгоритма генерации форм слов является получение словоформы F, соответствующей основе S и грамматическому значению G. Результат генерации форм слов всегда однозначен.

Алгоритм генерации состоит из двух этапов. На первом этапе производится поиск необходимой цепочки преобразований R по входным данным:

основе S и грамматическому значению G. Второй этап заключается в применении найденной цепочки R к основе S для получения требуемой формы F.

Алгоритм определения форм слов определяет грамматическое значение G и основу S, соответствующие исходной форме F.

Алгоритм определения форм слов заключается в переборе n цепочек преобразований в цикле. На каждой итерации извлекаются очередная цепочка Ri и соответствующие ей тип основы Ti и грамматическое значение Gi. Цепочка Ri', обратная цепочке Ri, применяется к входной словоформе F для получения основы Si. В зависимости от результата применения цепочки к строке, цепочки можно разделить на три вида:

1. целевые цепочки, приводящие к нахождению основы Si, тип которой соответствует типу основы цепочки преобразований;
2. ложные цепочки, приводящие к основе Si, тип которой не соответствует типу основы цепочки преобразований;
3. тупиковые цепочки, которые прерываются из-за неприменимости преобразований к форме.

Если найденной основе Si соответствует тот тип основы Ti, что и очередной цепочке, то есть цепочка Ri' – целевая, то результат, состоящий из полученной основы Si и грамматического значения Gi, добавляется к полученным результатам. На этом итерация заканчивается и происходит переход к следующей итерации. Число итераций цикла равно числу цепочек преобразований n. Перебор цепочек преобразований не заканчивается после нахождения первой применимой к форме F цепочки из-за существования в естественных языках омонимии. Поэтому не одна, а несколько цепочек преобразований могут быть применимы к форме F. При этом только одно решение является правильным.

Алгоритмы генерации и определения не зависят от конкретного вида преобразований. Цепочки, состоящие из преобразований, реализуются вспомогательными алгоритмами. Такое разделение позволяет сделать алгоритмы генерации и определения независимыми от конкретного естественного языка и присущих ему способов формообразования.

Решение интеллектуальных задач можно представить в виде графа в пространстве состояний. Вершинами графа являются состояния, а дугами – операторы. Корневая вершина является начальной ситуацией, а конечные вершины – целевыми или тупиковыми. Целевые вершины приводят к решению задачи, а тупиковые вершины – нет. Решение задачи можно рассматривать как последовательность операторов, которая преобразует начальное состояние в целевое.

Для поиска в графе целевых вершин существует два простейших метода поиска в графах: поиск в глубину и поиск в ширину. Данные методы поиска предполагают последовательный перебор возможных альтернатив. Чтобы повысить эффективность поиска, вводится правило выбора очередной вершины или отсекаются заведомо тупиковые направления поиска, не приводящие к решению. Такой подход называется эвристическим поиском.

Возможно представить решение задач генерации и определения форм слов как поиск в ориентированном графе в пространстве словоформ.

Построение графа в пространстве словоформ основано на следующем соответствии терминологии теории графов и метода генерации и определения форм слов:

- вершины графа – это словоформы, при этом корневая вершина – основа, промежуточные вершины – результаты применения преобразований цепочки к основе, а конечные вершины – словоформы парадигмы;

- дуги графа – это преобразования, которые необходимо произвести над одной формой, чтобы получить другую;

- путь (маршрут) в графе – это цепочка преобразований; путь от корневой вершины к конечной – прямая цепочка преобразований, путь от конечной вершины к корневой – обратная цепочка преобразований;

можно выделить два типа путей:

1. целевой путь – целевая или ложная цепочки преобразований;
2. тупиковый путь – тупиковая цепочка преобразований.

Пути, имеющие одну и ту же корневую вершину, составляют граф, и корневая вершина становится общей для них.

Оператор перехода из одного состояния в другое состоит из двух действий:

1. применить преобразование к текущей словоформе;
2. проверить условие перехода в следующее состояние: если преобразование применимо к форме, то перейти в следующее состояние, иначе считать данный путь тупиковым.

За счет отсечения тупиковых путей сокращается количество вершин, рассмотренных во время решения задачи.

В зависимости от задачи генерации и определения дуги в графе ориентированы в противоположные стороны. В случае задачи генерации дуги ориентированы от корневой вершины к конечным, а в случае задачи определения – в обратном направлении: от конечных вершин к корневой.

В пространстве словоформ задача генерации представляет собой поиск пути от корневой вершины к конечной, а задача определения заключается в нахождении такой конечной вершины, которая приведет к корневой вершине.